DOCUMENT RESUME

ED 458 244 TM 033 405

AUTHOR Lee, Yong-Won
TITLE Examining Passage-Related Local Item Dependence (LID) Using Q3 Indices in an EFL Reading Comprehension Test.
PUB DATE 2000-04-24
NOTE 69p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Factor Analysis; Foreign Countries; *High School Students; High Schools; Reading Comprehension; *Reading Tests; *Test Items
IDENTIFIERS *Item Dependence; Korea; Passage Dependency; *Q Index; Testlets

ABSTRACT

This paper reports the results of an analysis of a reading comprehension test using the Q subscript 3 (Q3) statistics developed by W. Yen (1984). Yen's Q3 can be a useful tool for examining local item dependence in the context of a reading comprehension test ir which a set of related items is followed by a reading passage. Q3 is basically a pairwise index of local dependence, an interitem correlation of residuals that are obtained by subtracting theta-predicted scores from raw scores. This study reports the results of a Q3 analysis of a 40-item reading comprehension test taken by 1,857 English-as-a-Foreign Language high school students in Korea. Average interitem correlations of raw score and residuals within "passages" and "item types" were computed and compared to the grand mean of Q3 among all item pairs. Factor analysis were also conducted both on matrices of interitem correlations of raw scores and residuals. A considerable amount of local dependence was found to exist among items within passages, but no sizeable evidence of local dependence was found among items within item types. Factor analysis also confirmed this finding by successfully identifying passage-related factors in residuals. These results provide support for grouping items based on the passages for testlet analysis. (Contains 9 tables, 5 figures, and 64 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 244

Examining Passage-Related Local Item Dependence (LID) Using Q_3 Indices in an EFL Reading Comprehension Test

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

Y.-W. Lee

2A TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee

TOEFL/Statistical Analysis/Assessment Division

Educational Testing Service

Rosedale Road Princeton

NJ 08541

Phone: 609-683-2401 (Office)

609-558-1331 (Home)

Fax: (609)-683-2130

Email: ylee@ets.org

TM033405

Paper presented at the annual conference of American Educational Research Association (AERA) in Division D in New Orleans, Louisiana, on April 24, 2000.

BEST COPY AVAILABLE

Abstract

In reading comprehension tests, it is quite common that a set of related items is followed by a reading passage, but such a cluster of items is very likely to be locally-dependent. From a psychometric standpoint, however, local dependence is not necessarily a matter of an *a priori* assumption, but more of an empirical phenomenon. We are not really sure whether the passage is the single most dominant factor that makes items locally dependent in reading comprehension tests, until we actually put it to an empirical test.

Yen's (1984) Q_3 can be a useful tool for empirically examining local item dependence in such a context. It is basically a pairwise index of local dependence, more precisely, an interitem correlation of residuals that are obtained by subtracting theta-predicted scores from raw scores. This study reports the result of Q_3 analysis of a 40-item reading comprehension test taken by 1857 EFL (English as a Foreign Language) high school students in Korea. Average interitem correlations of raw score and residuals (Q_3) within "passages" and within "item types" were computed and compared to the grand mean of Q_3 among all item pairs. Factor analyses were also conducted both on matrices of interitem correlations of raw scores (*Phi* and *Tetrachoric*) and residuals (Q_3).

A considerable amount of local dependence was found to exist among items within passages, but no sizable evidence of local dependence was found among items within item types. Factor analysis also confirmed this finding by successfully identifying passage- related factors in residuals. These results provide support for grouping items based on the passages for testlet analysis.

Examining Passage-Related Local Item Dependence (LID) Using Q_3 Indices in an EFL Reading Comprehension Test

I. Introduction

In English as a Foreign/Second Language (EFL/ESL) reading comprehension tests, it is very common that a series of related items are based on the same reading passage. These items can be placed after a passage, as in a traditional multiple-choice test, or embedded in the passage itself, as in a *cloze* test. When multiple items are used for a single passage, it could constitute a viable test format. Such a format makes it possible to measure examinees' understanding of the material from various perspectives and is cost-effective both for item writers and examinees. That is, the item writers would not have to spend much time preparing the passage to generate a single question from it, and the examinees also would not need to read a long passage to answer only one question.

Despite such strengths, however, the passage-based item sets pose a problem for test analysis and construction within the framework of contemporary measurement models. It is assumed implicitly in classical measurement theory and explicitly in item response theory (IRT) that test items should be *locally independent* (Henning, 1989; Sireci, Thissen, & Wainer, 1991; Yen, 1984, 1993). When a series of items share a common stimulus material, it is very likely that the local independence assumption is violated among those items. In consequence, biased estimation of examinee abilities, item information, and test reliabilities can be obtained, when psychometric models assuming local item independence (especially, IRT) is applied to the tests that include

sets of locally-dependent items (Ackerman, 1987; Ferrara, Huynh & Michaels, 1999; Wainer & Thissen, 1996; Yen, 1993).

One intriguing solution to the problem is to group together items that are (or likely to be) locally-dependent and treat them either as a single polytomous item or as a cluster of items sharing an additional item parameter. The first approach, which is sometimes called the *testlet approach*, involves scoring the superitem as a whole and fitting various polytomous measurement models to the test (Andrich, 1985; Lee, 1998; Rosenbaum, 1988; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wilson & Iventosch, 1988). In this approach, however, base-level items are no longer identifiable after scoring in the analysis. On the other hand, the second approach that attempts to fit the *testlet response model* recognizes the item cluster as a superordinate unit of analysis, but does not try to score them as a single superitem. An additional item parameter is estimated, instead, to reflect local dependence within the item set in the context of the two- or three parameter logistic IRT model, which is assumed to be common across all the items within the item set (Bradlow, Wainer & Wang, 1998; Du, 1998; Wainer, Bradlow & Du, 1999).

These approaches become most effective, when the item grouping principle for locally-dependent items is determined *a priori* and when local dependence is severe enough to produce biased estimates of item and person parameters in the test. As a matter of course, passage-based item sets are the most likely candidates for such item clusters in reading comprehension tests. Local dependence, however, is not necessarily a matter of *a priori* assumption, but more of empirical phenomenon (Steinberg & Thissen,

1996). In other words, we cannot be really sure whether passages are the single dominant factor that makes items locally dependent and how serious the level of local dependence among items are, until we actually put it to an empirical test. Even worse, if there are mutually non-exclusive, overlapping characteristics (*e.g.*, passages and item types) that equally contribute to local item dependence, it might not be very efficient to group items according to a single criterion (*i.e.*, the sharing of stimulus material) in order to minimize its impact.

Furthermore, the issue of local dependence is also closely related to test dimensionality. If a cluster of items is found to be locally-dependent, it may indicate that there is a secondary dimension that is measured by these items in addition to the primary one measured by the test as a whole. If this secondary dimension is also linked to an important dimension that is of relevance to the intended construct (s), it may not be appropriate just try to minimize the impact of local item dependence or absorb its impact into an additional item parameter in the unidimensional IRT framework, as in the two approaches described above. Multidimensional approaches, if feasible models exists that can handle various secondary dimensions simultaneously, might be more appropriate in such cases.

This paper basically addresses the very issue of local dependence in EFL reading comprehension tests. It is aimed at empirically examining whether passages and item types can be shown to be equally contributing to local dependence in an EFL reading comprehension test by using Q_3 statistics and factor analytic methods based on Q_3. To that end, the author first discusses local dependence and Q_3 statistic in relation to test

dimensionality; second, identifies two possible sources of local dependence in reading comprehension tests; third, presents the results of interitem correlation analyses of raw scores for 10 passage-based item sets and 4 item-type-based item sets; fourth, report the Q_3 analyses of 10 passage-based item sets and 4 item-type-based item sets; finally presents the results of factor analyses based both on matrices of interitem correlations of raw scores (*phi* and *tetrachoric*) and residuals (Q_3).

1 Local dependence (LD), test dimensionality, and LD indices

Local dependence refers to a lack of local independence (LI) among items. In classical measurement theory, it is assumed that errors of measurement are not only statistically independent of true scores (Suen, 1990) but also statistically independent among different forms or items (Cureton, 1965). In the case of α-coefficients for internal consistency, errors of measurement are assumed to be random and not be correlated across items. If measurement error is positively correlated over items, the items are not statistically independent of each other (Yen, 1993). In IRT, the assumption of local independence stipulates that responses to two different items in a test should be statistically independent of each other for examinees at a fixed ability/trait level. Expressed differently, an examinee's response to one item should not influence his or her response to other items in the test (Hambleton, & Swaminathan, 1985). If the content of one item provides clues to the answer of another test item, the two items are not independent, and the assumption is violated. That is, they are locally-dependent.

Technically, local independence can be defined as a zero correlation between the two items for individuals with the same ability level, especially from the IRT perspective

(Henning, 1989; Weiss & Yoes, 1991). It does not mean, however, that the two items should be uncorrelated along the entire ability range (as is often misunderstood), but does mean that they should be correlated only through the latent trait (θ) the test is measuring (Henning, 1989; Lord & Novick, 1968). In IRT, such a correlation between the two items is expressed as a function only of the model's item and person parameters (Wilson, 1988). In this sense, the definition of LI can be further refined in factor analysis terminology as a zero correlation between the two items, after the impact of the measurement construct (θ) is partialled out from the item scores. When there is still a significant level of a correlation remaining between the residuals of the two items, then the two items are considered to be locally-dependent. If the two items are locally-dependent, it may potentially indicate that there is a secondary dimension (s) that is tapped by these items, but unaccounted for by the unidimensional IRT theta (θ).

Due to the equivalence between the assumptions of local independence and unidimensionality in the unidimensional IRT, factor analytic techniques based on raw scores can be used as a rough check on the extent to which a set of test items satisfies the assumption of local independence at an aggregate level in the test (Hambleton & Swaminathan, 1985; Spray & Ackerman, 1987). Yen (1984) presented three different kinds of fit statistics as measures of local dependence at an item, item pair, and aggregate levels: Q_1, Q_2, and Q_3. First, Q_1 refers to the traditional measures of item fit in IRT that typically involve comparisons between observed and predicted item characteristic curves using a chi-square test (Yen, 1981). Second, the Q_2 statistic (Wollenberg, 1982) was designed to be sensitive to local dependence at an aggregate level (over items) for the

Rasch model; Yen (1984) extended its application for the three-parameter logistic (3PL) model and suggested the addition of a sign (+ or -) to indicate the direction of local dependence. Third, Q_3 is basically a pairwise index of local dependence, more precisely, a correlation of item residuals after the theta-predicted scores are subtracted from the raw scores.

Recently, Thissen and Chen (1993) also proposed four more IRT-based LD indices: the Pearson's χ^2, the Likelihood Ratio G^2, the Standardized ϕ–coefficient Difference, and the Standardized Log Odds-Ratio Difference. These four statistics are somewhat similar to Q_3 in the sense that they test item covariation that is not accounted for by the item response theory model, but do not require the point estimate of theta for all examinees because they rather use the observed and expected frequencies of item pairs. Chen and Thissen (1997) have shown that, as compared to Q_3, the first two indices are somewhat less powerful for "the underlying LD" stemming from the contents of items, but equally powerful for "the surface LD" that often found in a series of not-reached items at the end of the test.

Among a number of these local dependence indices explained so far, Q_3 was selected in this study, because this index is conceptually most consistent with the definition of LD as "a zero correlation between two items after the theta is partialled out". Q_3 is a pairwise index of item dependence that was developed by Yen (1984) based on ideas suggested by Kingston and Dorans (1982). Kinston and Dorans (1984) initially suggested on computing a correlation of item scores and partialling out the impact of theta from the correlation, but it is known that this statistics is only capable of removing a

linear relationship between item scores and thetas, when there is a non-linear logistic relationship between them. For this reason, Yen (1984) proposed, as an alternative to this, that we could remove the impact of theta from the item scores first and compute a correlation between the item residuals after that. That is,

$$d_{ik} = u_{ik} - \hat{P}_i(\hat{\theta}_k) \ , \quad and$$
$$Q_{3_{ij}} = r_{d_i d_j}$$

where, u_{ik} is the score of the kth test taker on the ith item. The computation of Q_3 requires a point estimate, θ, for each examinee. Because d_{ik} and d_{jk} are random errors, we may assume that, when the three-parameter model is true, they will be distributed approximately as bivariate normal variables with a zero-correlation.

In reality, Q_3 statistics between a pair of items may tend to be slightly negative, even when the two items are actually locally-independent, because raw item scores are involved in the calculation of *theta* (θ) estimates (Yen, 1984). To put it another way, even though there is no local dependence between the two items, the Q_3 could be slightly negative due to the uncorrected part-whole contamination between item scores and *theta* (θ) estimates, which may be analogous to uncorrected point-biserial correlations between item scores and total scores in some senses. It may be also possible to conduct a factor analysis on a matrix of residuals (Q_3) after the theta is partialled out to investigate the factor structure of residuals that may sheds a light on systematic sources of local dependence.

2 Passages and item types as possible sources of local item dependence

In the context of reading comprehension tests, of course, the most probable source of local dependence might be the stimulus material (*i.e.*, a passage) shared among a group of related items. Nevertheless, it may not be reasonable to completely exclude the possibility that there might be other equally influential factors than the sharing of stimulus material that cause local item dependence in reading comprehension tests (Rosenbaum, 1988; Wilson & Adams, 1995; Yen, 1993). "Item types," which are often believed to tap different levels or aspects of reading comprehension, were selected as a potentially competing factor for local item dependence that can overlap with the sharing of passages among items. Reasons for choosing these two factors are discussed in terms of local dependence and test dimensionality hereafter.

Passages. The interaction of text content with readers' background knowledge in reading performance has long been documented in both reading theory and research. Schema theory, for instance, has suggested that readers who are familiar with a particular topic develop a base of knowledge that facilitates acquisition of further information on that topic (Carrell, 1987; Carrell & Eisterhold, 1983; Rumelhart, 1980). Its implication for second language reading assessment is that examinees' performance on reading comprehension tasks might be influenced by their prior knowledge on the topic as well as their reading comprehension ability in the target language (Clapham, 1996). Such insights have stimulated numerous studies in differential item functioning (DIF) in EFL/ESL reading comprehension tests, which is attributable to differences in examinees'

areas of specialization (Alderson & Urquhart, 1985; Brown, 1982; Erickson & Molloy, 1983; Hale, 1988; Henning, 1990).

Another perspective useful in understanding passage-related local dependence is to see, at some risk of oversimplification, the construct "reading comprehension" measured in the test as a latent dimension underlying two layers of data, that is, (1) examinees' scored responses to each item within a passage and (2) a sum of item scores in the passage. In IRT, the inference is made of the measurement construct from scored responses (e.g., right or wrong) to individual items that are sampled from the universe of possible items in the target domain (Anderson, 1972), which are assumed to be statistically independent of each other. In reading comprehension tests, however, all items within a single passage are aimed at assessing examinees' comprehension of the same text from different perspectives, and thus an examinee's performance on one item is likely to affect their performance on another item within the passage. Ironically, what the test as a whole intends to measure is not usually the comprehension of a particular text *per se*, but rather general "reading comprehension" ability across multiple texts on various topics included in the test. Therefore, unique item dependence related to the text unaccounted by this general latent trait could lead to local dependence in an IRT sense.

For this reason, if only one item is given from each distinct passage in the test, passage-related local dependence may not become a serious issue, even though the item might still be subject to DIF examination. The interaction between the content of a specific passage and examinees' prior knowledge may be isolated on a single item in such a case. Therefore, the unique item variance unexplained by the common factor would be

less likely to lead to significant correlated errors among adjacent items, because they are all based on different text contents. When a passage is shared by a group of related items, however, some extra covariations are more likely to be present among some pairs of these items due to the interaction between the text content and examinees' prior knowledge that is unaccounted for by the target trait (θ) measured by the test as a whole. In other words, the shared passage can potentially add a secondary dimension to test dimensionality.

Thissen and others (1989) have roughly demonstrated the existence of passage-related factors in a four-passage, 22-item test of reading comprehension through the full information factor analysis. Schedl, Gordon, Carey, and Tang (1996) also found out through dimensionality analysis and multidimensional calibrations that there might be a minor secondary factor, possibly related to passage content or passage position effects in the reading section of TOEFL. However, it should be noted that local dependence is defined in this study in terms of the relationship between the two items in relation to the unidimensional latent trait (θ) measured in a test as a whole. Moreover, the undimensional trait defined in IRT may not necessarily mean a single latent trait in the strictest sense, but also could be the same composite of abilities, as shown by Reckase, Ackerman, and Carlson (1988).

Then, an interesting empirical question comes: Is it empirically demonstrable that some passage-related factors are not absorbed completely into the undimensional latent trait (θ) modeled by IRT and become the causes of local dependence among some items sharing the passages in the EFL context? If so, are they single dominant factors contributing to local dependence consistently throughout the test?

Item types. "Item type" can mean different things in reading assessment depending upon what aspects of test tasks we are focusing on. It may mean different response modes for reading assessment tasks including true/false, matching, multiple-choice, short-answer items, and essays (Savignon, 1983), different relationships between the questions and the information in the text necessary to answer the question (Pearson & Johnson, 1978) or different reading subskills required to complete the task (Barrett, 1976; Munby, 1978).

Rosenbaum's (1988) statement has provided a starting point for exploring item types as an additional factor of local item dependence that can be equally influential and potentially overlapping with the shared reading text. He has basically point out in the discussion of the IRT models for locally-dependent items that:

> It is entirely possible, however, that items will exhibit dependence without explicitly sharing material. There are many types of items which seem difficult when first attempted, but which seem to become somewhat easier with practice on similar items. Possibly, a few cognitive tasks are specific to all items of this type, and "practice" involves completing those few common tasks and storing the results for use on similar items. Certainly, one can construct mathematical word problems or verbal analogies that are so parallel in nature that the sharing of cognitive task is almost undeniable. Whether this happens to a noticeable degree with the types of items normally included in tests is not, at present, known.... (p. 358-359).

His main intention seems to be to bring more attention to those items among which a few common cognitive tasks are shared, but it is not clear whether these common tasks would be something that is construct-relevant or construct-irrelevant or combination of both. If all these common cognitive operations can be explained by the underlying latent trait

measured in the test as a whole, then it would be unlikely to lead to local dependence, as defined in this study: no correlation after the impact of the latent trait is removed. In order for these common cognitive tasks to be related to local dependence, these common cognitive tasks should be somewhat construct-irrelevant, or should be somewhat related to test method effects, if Bachman's (1990) term is borrowed. In this sense, Wilson and Adams (1995) rather point to a group of item that share a common response format as a plausible candidate for an item cluster that can overlap with the sharing of stimulus material in the test.

In the context of reading comprehension, the discussion of item types based on subskills seems to be very intriguing from this perspective. Various "item types" are often used in reading comprehension tests to cover a wide range of reading comprehension activities. Such a practice seems to have been linked to the development of various taxonomies that have been proposed to represent different levels, aspects, or subskills of reading comprehension (Barrett, 1976; Bloom, 1956; Davies, 1968; Gray, 1960; Mossenson, Hill & Masters, 1987; Munby, 1978; Pearson & Johnson, 1978) in the syllabus design and test construction. Some of them were developed speculatively through the content expert's intuition and introspection (Munby, 1978), and others were empirically derived using factor analytic techniques (Davis, 1968) or employing the Rasch content-referencing technique (Mossenson *et al*, 1987). Even in the EFL/ESL setting, such a taxonomic approach has been widely used for the development of test content specification and test items (Alderson, 1990a; Carroll, 1968; Lumley, 1993).

However, the existence of these reading subskills has been an issue of heated

debate among language testers, as is evidenced by the recent debate between Lumley (1993,1995) and Alderson (1990a, 1990b, 1995). Thus, item types defined by these subskills can or cannot be related to local dependence, depending upon how reading subskills are conceptualized. First, the argument has been argued that the reading process is more complex than is implied by the these discrete lists of skills and it may not be possible to identify and isolate a single dominant subskills in the test items (Alderson, 1990a, 1995; Alderson & Lukmani, 1989). From this perspective, it is not worthwhile to link local dependence to item types at all.

Second, it has been believed that a single dominant skill is identifiable through the judgmental methods in many items, without which it might not be possible to answer the items. The related idea is that the hierarchy from the lower- and higher-order subskills can be verified in the item difficulty estimates of the test items in the unidimensional measurement framework (Lumely, 1993). From this viewpoint, the issue of local dependence may not arise seriously, because the item difficulty parameter is defined along the unidimensional scale. However, it would be still worthwhile to empirically examine the issue.

Third, the debate on the subskills is often focused on the divisibility of comprehension subskills (Johnston, 1983; Davis, 1944,1968), the qualitatively different dimensions for the lower-to-higher order subskill on the hierarchy, or the multidimensionality of the test (Buck, 1994). Buck (1994) argues that second language comprehension involves multidimensional interactions of many different linguistic and nonlinguistic skills and processes and that unidimensional measurement models, such as

IRT, are not appropriate for second language assessment. Under such circumstances, the item-type-related local dependence needs to be carefully examined.

Then, an intriguing question would be whether or not it is possible that common cognitive attributes (linguistic or non-linguistic) shared among items in the same item type can make all items sharing those cognitive features locally-dependent due to the minor factors that are not explained by the undimensional IRT theta or due to the practice effects over similar items in a test, as has been pointed out by Rosenbaum (1988).

II. Method

1. Sample

Examinees were tenth and eleventh-grade *English as a Foreign Language* (EFL) high school students enrolled in five different high schools in a metropolitan area including Seoul (the capital city of South Korea) during Fall 1997. The total number of examinees was 1857. There were 613 tenth graders and 1244 eleventh graders. The number of participants at each school ranged from 304 to 413.

2. Instrument

After reviewing the literature (Barrett, 1976; Munby, 1978; Pearson & Johnson, 1978), four item types were chosen: (1) recall of details, (2) main topic inference, (3) inference of details, and (4) prediction of content (See Table 1). These four item types were arranged to make them closely correspond to the different relationships between the questions and the information in the text required to answer the question (Pearson & Johnson, 1978): textually explicit, textually implicit, and scriptally implicit. Besides, the wordings were used for the item stems of the same type of items.

Insert Table 1 about here

The instrument used in this study was a 40-item multiple-choice reading comprehension test that was developed for Korean high school students by the author. The test was pilot-tested, revised, and reviewed by experts with relevant background, before it was used for the current program of research. Ten reading passages were used, and two reading passages were selected from each of the five subject areas of humanities, social sciences, natural sciences, engineering, and arts/sports (See Table 2).

The length of passages ranged from 148 to 203 words when contracted forms (*e.g.*, don't) and compound words (*e.g.*, Anglo-Saxon) were counted as single words. Some of the passages underwent a slight revision in order to create the same 4 item types across all of the 10 passages used in the test. Even though there are no established rules regarding how many is the most appropriate number for comprehension items to accompany a single passage, it was helpful to look at Harris's (1969) suggestion that, when a passage ranges from 100 to 250 words, it can be followed by 4 to 7 items.

Insert Table 2 about here

As shown in Table 2, four multiple-choice items were constructed for each passage, and those four items were of the same item format and arranged in the same sequence throughout all 10 passages: (a) The first item following each passage asked examinees what the main topic of the passage was; (b) the second, what was mentioned or not mentioned in the passage as a recall of factual details; (c) the third, which phrase should follow a preceding phrase in the underlined part as a factual inference of details (*e.g.*, A is B because _____ .); and (d) finally the fourth item, what kind of content would follow this passage as a prediction of content. The content of the test was originally developed in English, and then the instructions, stems for the first, second, and fourth questions, and the choice options for the fourth questions were translated into Korean in order to help examinees understand the questions easily. The accuracy of translation was checked by three bilingual graduate students studying TESL (Teaching English as a Second Language), who are proficient both in Korean and English.

Five different forms (or spirals) of the same test were prepared by changing the order of passages in the test. The internal consistency reliability (KR-20) for the 40-item EFL reading comprehension test was .89. The marginal reliability for the same test estimated by fitting the 3-parameter IRT model was .90. The scale mean and standard deviation for the total score were approximately 20.4 and 8.58 (See the Results section to see how reliability changes when item clusters were used as scoring units).

3. Procedures

Students' participation was solicited by cooperating EFL teachers at each school. Five different test forms were used at each school to minimize the chance of cheating

among students. The cooperating teachers at each institute distributed five different forms randomly among students within each school, and the students took the test under the supervision of the cooperating teacher. They were given 60 minutes to answer 40 questions on the test. They were instructed to mark the answers directly on the test sheet.

4. Data Analysis

Data information on the test sheets were first keyed into the computer program Microsoft EXCEL 5.0 and then imported into the SPSS Windows 6.0 program for recoding, scoring, and reliability analysis. In order to ascertain whether the decreased reliability estimates are obtained when passages are used as the unit of scoring, item scores within each passage were totaled to 10 individual passage scores. Likewise, 40 item scores were also aggregated into 4 item-bundle scores for the same item type, test reliability estimates were obtained.

a Average interitem correlations of raw scores: A matrix of 780 non-redundant coefficients of interitem correlation was obtained for each of the 3 different methods of computations: (1) phi-coefficients for listwise deletion; (2) phi-coefficient for pairwise deletion; and (3) tetrachoric coefficients for listwise deletion. The SPSS Windows program was used to compute phi-coefficients, after missing cases were deleted listwise or pairwise for analysis, as was described in the literature (Little & Rubin, 1987). Tetrachoric coefficients were also computed by using the DIMTEST program (Stout, Douglas, Junker & Roussos, 993), because there were some disadvantages of phi-correlations mentioned in the literature in connection to the artificial difficulty factor in dimensionality analysis (Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons,

1983). The correlational matrices produced by SPSS Windows and DIMTEST were edited on a text-editor program and were imported into the computer program Microsoft Excel (5.0). The average interitem correlations were calculated across all item pairs, within passages, and within the same item types across passages. The Fisher-Z and inverse Fisher-Z transformations were conducted respectively before and after the all the correlation coefficients were added and averaged.

b Average interitem correlations of residuals (Q_3): MULTILOG 6.2 (Thissen, 1991) and IRTNEW (Chen, 1996) were used together to obtain the interitem correlations of residuals (Q_3). For a manual computation of Q_3, the MULTILOG and EXCEL programs were used together as well. MULTILOG is an IRT computer program that can handle binary and multiple category data and allow for imposing constraints on the parameters. IRTNEW is a specialized computer program designed to compute indices of local dependence for binary item responses using the item parameter output from major IRT programs including MULTILOG.

First, MULTILOG 6.2 was used to obtain the item parameter estimates for the three-parameter logistic model (3PL). The MML (Marginal Maximum Likelihood) estimation procedure was employed to estimate the item parameters. To maximize the accuracy of item parameter estimates, the number of iterations was set at 100, which is much higher than the default value, 25. The quadrature points were also set to be narrower than originally set within the program by default, with the interval of .2 from -4.0 to 4.0. For the two data sets with missing cases deleted listwise and replaced by a missing code of 9, the iterations converged within the 50 cycles of iterations.

The computer program IRTNEW (Chen, 1996) was run on the data file and the MULTILOG output file of item parameters. Based on this information, IRTNEW estimated examinee abilities (θ), computed residuals between observed raw scores and *theta* (θ)-predicted scores, and finally produced a matrix of Q_3. Both MLE (Maximum Likelihood Estimate) and MAP (Maximum *A Posteriori*) estimates of thetas were used in computing Q_3.

Second, the manual computation of Q_3 also had to be conducted for the second data set in which missing cases were replaced by a missing data code of 9. IRTNEW performed well for the data set with missing cases deleted listwise (N=1072) and MAP estimation, but it did not perform well for the data set containing missing data points, especially when MLE estimation was used. IRTNEW just stopped running without completing the necessary steps when encountering a non-converged *theta* (θ).

MLE estimates of item parameters from the initial random run of MULTILOG and estimates of *theta*s from the second scoring run of MULTILOG were imported onto EXCEL for the second data set; *theta* (θ)-predicted scores were computed and subtracted from raw scores, and the score residuals were obtained. Then the score residuals were imported into the SPSS Windows, and a matrix of Q_3 finally was produced. Sixteen cases with no *theta* estimators and sixty cases with non-converged *theta*s were deleted in the computations of Q_3.

The Q_3 correlational matrices obtained either through IRTNEW and manual computations were edited and imported into Excel (5.0) in order to obtain the coefficients

of average interitem correlation of residuals across all items and within and between passages.

c Principal component analysis (PCA) and factor analysis (FA): Principal component analyses were conducted on both the raw and residual scores of the 40 items by using the SPSS Windows 6.0 program. Both PCA and FA basically use correlational matrices from scores, that is, matrices of *phi-* and *tetrachoric* correlations for raw scores and of *Pearson* correlation correlations for residual scores (Q_3). Since the computation of *tetrachoric* correlations was not possible in SPSS, the *tetrachoric* correlations were obtained by using DIMTEST and the correlational matrix was imported into SPSS and used for PCA instead of raw scores themselves. DIMTEST accepts only binary data and automatically treats missing values as incorrect responses. For this reason, only the data set with missing values deleted listwise (N=1072) was used in the analysis. Eigenvalues were obtained for principal components extracted and were plotted. The scree plots of the extracted components were also obtained both for raw and residual scores and compared.

For residual scores, the varimax rotations were done on extracted components to make factor loading more interpretable. After the appropriate factor solutions were obtained, factor loadings of 40 items were plotted in the 3-dimensional space of 3 selected factors to present visually the relationships between items and factors.

III. Results

1 Test reliability

The *alpha*-coefficient was lowered from .89 to .87 as expected, when the 40 item scores were aggregated into 10 passage-scores, as shown in Table 3. Nevertheless, the mean score (20.4) and standard deviation (8.6) remained the same, and the average correlation among items increased from .17 (dichotomous) to .40 (polytomous). Sireci, Thissen, and Wainer (1991) and Wainer and Thissen (1996) explain that the first α-coefficient (.89) would be an overestimate of the true α-coefficient (.87), and that the difference between the two is the inflated portion of α-coefficient due to the local dependence among items within the passages.

Insert Table 3 about here

Interestingly enough, aggregating dichotomous item scores into a smaller number of item-bundle scores does not always decrease the value of reliability estimate of the same test. For example, the alpha-coefficient virtually remained the same, when the 40-item scores were summed into 4 item-type testlet scores. Actually, the coefficient increased insignificantly from .8929 to .8931. The mean and standard deviation of the total scores remained the same again, but the average correlation among items (or testlets) increased from .17 (dichotomous) to .68 (polytomous). As long as Sireci *et al.*'s (1991) claim is valid, such a tendency could be seen as the first overall indication that no significant local dependence exists with item types in the test.

2 Interitem correlations of raw scores within passages and item types

The grand mean of the 780 interitem correlation coefficients, the mean of 60 within-passage correlations, the mean of 180 correlations between the same types of items (within item-types) across passages, and the mean of 540 correlations between different types of items across passages were computed for each of the 3 different correlation matrices via Fisher-Z and inverse Fisher-$\bar{Z}$ transformations. The three different correlation matrices were from: (1) *phi*-coefficients for the listwise deletion; (2) *phi*-coefficients for the pairwise deletion; and (3) *tetrachoric* coefficients for the listwise deletion. The results of these computations are summarized in Table 4.

Insert Table 4 about here

The sizes of differences were computed between the average within-passage interitem correlations and the average within-item-type interitem correlations. Likewise, the difference size was also computed between the average within-passage interitem correlations and the grand means. Their values are also shown in parentheses in Table 4.

a Average correlations within passages and item types as a whole: First, the grand means of *phi*-coefficients across all item pairs were .173 for the listwise deletion and .166 for the pairwise deletion. The *tetrachoric* correlations were higher than the *phi*-correlations in general across all item pairs. The grand mean of *tetrachoric* coefficients was .276 for the listwise deletion.

Second, the mean of within-passage correlations across 10 passages was higher than the grand mean of all item pairs in all three different methods. The average *phi*-correlations were .229 for the listwise deletion and .215 for the pairwise deletion, but the average *tetrachoric*-correlation was .363 for the listwise deletion. The difference between the means of the within-passage interitem correlations and the grand means were .058, .050, and .097 for the above three methods (See Table 4).

Third, the mean of within-item-type correlations across 4 item types was either equal to or slightly lower than the grand mean of all item pairs in all 3 methods. The average *phi*-correlations between the same types of items were .173 and .166 for the listwise and the pairwise deletion methods, but the average *tetrachoric* correlation was .275 for the listwise deletion method.

Last, the mean of within-passage correlations across 10 passages was higher than the mean of within-item-type correlations across 4 item types in all three methods. The difference between the average within-passage interitem correlations and the average within-item-type correlations were .058 for the *phi*-listwise deletion and .050 for the *phi*-pairwise deletion method, but .098 for the *tetrachoric*-listwise deletion method.

b Average correlations within each passage and item type. In order to ascertain whether the same patterns can be observed at the levels of each passage and each item type, the average interitem correlations within passages and within item types were examined passage by passage and item-type by item-type, respectively. These average correlations were also computed using the same three methods. The results of these computations are presented in Table 5.

Insert Table 5 about here

First, there were some variations among 10 passages in average interitem correlations, as in Table 5. In the *phi*-listwise method, for example, the average within-passage correlation ranged from .176 to .286, with a mean of .229. There were 5 passages that have lower average correlations than the mean, which were Passages 1 (Composition), 2 (English Vocabulary), 7 (Polygraph), 9 (Piano Tuning), and 10 (Roller Skating). The same patterns were also observed in the *phi*-pairwise and the *tetrachoric*-listwise deletion methods.

Second, among 4 item types, the first two (*i.e.*, main topic, recall of details) were comparatively high in the average interitem correlations, but Item Type 4 (prediction of content) was comparatively lower than these two item types. Item Type 3 (inference of details) was located near the mean of all 4 item-types in the average within-item-type correlations. In the *phi*-listwise deletion method, for example, the average within-item-type correlations for Item Types 1, 2, and 3 were .190, .199, and .174, which were higher than the mean of the total item-type correlations (.173). However, the average within-item type correlations for Item Type 4 was .127, which was lower than the mean. The similar patterns were also found in the computation results from the other two methods.

Third, most interestingly, the average within-item-type interitem correlations for the top two item types in average correlations were higher than the average within-passage interitem correlations for the bottom three passages in average correlations. For

instance, the average interitem correlation for Item Type 2 (recall of details) was the highest (.199) among 4 item types in the *phi*-listwise deletion method, which was even greater than the average within-passage interitem correlations for Passages 1 (.178), 2 (.176), 9 (.182), and 10 (.193). Item Type 2 (main topic) also had higher average interitem correlations (.190) than Passages 1, 2, and 9. Except these cases, all other individual passages had higher average interitem correlations than all of the 4 item types. The same patterns were observed in other two methods.

In sum, comparisons of average interitem correlations of raw scores within passages and item types have provided sound evidence that passages create stronger dependence among items than item types at the levels of overall means. Nevertheless, only weaker partial evidence was obtained in the comparisons of the individual passages and item types.

3. Interitem correlations of residuals between raw scores and θ-predicted scores (Q_3)

Two different data sets were used for the estimation of item parameters: (1) the data set with missing values deleted listwise (N=1072) and (2) the data set with missing values recoded as a missing code of 9 (N=1857). When the thetas were estimated, both the MLE (Maximum Likelihood Estimate) and MAP (maximum *A Posteriori*) methods were used for each data set. In addition, when the missing code of 9 was used for unanswered items in the theta estimation, as in the second data set, either listwise or pairwise deletion had to be conducted in the final stage of Q_3 computations in both MLE and MAP estimations. The results of computations by these 6 different methods are shown in Table 6.

Insert Table 6 about here

The size of the differences were also computed between the average within-passage Q_3 and the average within-item-type Q_3 and between the average within-passage interitem Q_3 and the grand means of Q_3 among all item pairs, which were computed by means of the Fisher-Z and inverse Fisher-Z transformations. Their inverse Fisher-Z-transformed values are shown in Table 6.

a Average Q_3 within passages and item types: First, the grand means of 780 coefficients of interitem correlations of residuals (Q_3) over all item pairs were of negative values near zero, no matter what methods of theta estimation or of treating cases with missing values were used. The grand means of Q_3 among all item pairs were -.016 for both the MLE and the MAP methods, when the first data set (N=1072) was used. When the second data set (N=1857) was used, on the other hand, the grand means were -.018 both for the MLE/listwise and the MLE/pairwise deletion methods and -.016, and -.019 for the MAP estimation/listwise deletion and the MAP estimation/pairwise deletion methods, respectively.

Second, the overall means of 60 within-passage interitem correlations of residuals (Q_3) across 10 passages were all positive in direction and about twice bigger than the grand means in magnitude in all 6 different methods of analysis. The means of the within-passage interitem correlations of residuals were .051 for both the MLE and the

MAP estimation methods, when the first data set was used. When the second data set was used, the means were .051, .044, .052, and .042 for the MLE estimation/listwise deletion, the MLE estimation/pairwise deletion, the MAP estimation/listwise deletion, and the MAP estimation/pairwise deletion methods, respectively.

The differences between the average within-passage Q_3 and the grand mean of Q_3 were .067 for both the MLE and the MAP estimation methods in the first data set and .068, .062, .068, and .061 for the above four methods used in the second data set, respectively (See Table 6). Thus, the differences in average interitem correlations between the within-passage items and the within-item-type items became slightly bigger in residuals (Q_3) than in raw scores (*phi*).

Third, the overall means of 180 interitem correlations of residuals (Q_3) within item types were also of all negative values in 6 different methods of analysis. The average Q_3 within item types was -.019 for both the MLE and the MAP estimation methods for the first data set; when the second data set was used, however, they were -.020, -.020, -.018, and -.022 for the MLE estimation/listwise deletion, the MLE estimation/pairwise deletion, the MAP estimation/listwise deletion, and the MAP estimation/pairwise deletion methods, respectively. Such negative correlations near zero might be an indication of no significant positive local dependence between the same types of items located in different passages at an aggregate level, as Yen (1984) pointed out.

The differences between the average within-passage Q_3 and the average within-item-types were .069 for both the MLE and the MAP estimation methods, when the first data set was used (See Table 6). When the second data set was used, however, they were

.071, .064, .070, and .064 for the MLE estimation/listwise deletion, the MLE estimation/pairwise deletion, the MAP estimation/listwise deletion, and the MAP estimation/pairwise deletion methods, respectively.

b Average Q_3 for each passage and each item type: In order to find out whether the positive local item dependence can be observed consistently at each individual passage, the average within-passage Q_3's were examined passage by passage. Likewise, the average within-item-type Q_3's were also examined item-type by item-type, and each item type was compared to each passage in Q_3. Because average Q_3's based on MLE and MAP were identical in the first data set (N=1072) and very similar in the second data set (N=1857), computation results only by MLE estimation are reported here. For the second data set, both listwise and pairwise deletions were used in the final stage of Q_3 computation after theta estimation. The results from these three methods of computations are presented in Table 7.

Insert Table 7 about here

First, the most remarkable observation is, as shown in Table 7, that all of the 10 passages had positive values of average within-passage Q_3 in all three different MLE methods, even though there were some variations in their magnitude. When the MLE method was used for the first data set (N=1072), for example, the average within-passage Q_3's ranged from .023 to 076 across 10 passages, with a mean of .051. Five passages had lower average within-passage correlations than the mean, which were Passages 2 (English

Vocabulary), 3 (Means for Trade), 4 (Dream), 5 (Yogurt), and 9 (Piano Tuning). Especially, Passages 3 and 5 had the highest average within-passage interitem correlations in raw scores, but turned out to be the third and fourth from the bottom in Q_3. Passage 5 also had a greater average within-passage correlation than the mean, but turn out to have the lowest average within-passage Q_3. The similar patterns were observed in other two computation methods.

Second, contrary to the within-passage Q_3's, all of the 4 item types had negative average within-item-type Q_3 in all three methods of computation. When the MLE method was used for the first data set, for example, the average within-item-type Q_3's ranged from -.029 to -.007 for the four item types, with a mean of -.019. Previously in the average correlations of raw scores, Item Type 2 had higher positive average interitem correlations than P01, P02, P09, and P10, and Item Type 1 had higher positive average interitem correlations than the first three passages of these four. However, none of the four item types were higher Q_3 than any of the 10 passages.

In sum, comparisons of average Q_3 within passages and item types have shown that positive local dependence exists among items within passages, but no significant positive local dependence were found to exist within item types. Furthermore, the comparison of individual passages and item types in Q_3 provided more convincing pieces of evidence that passages are consistently dominant factors that contribute to positive local dependence among items throughout the test.

3. Factor analysis on the interitem correlation of raw scores and residuals

Principal component analyses (PCA) were first conducted both on raw scores and residuals, and their results were compared in order to see whether the test dimensionality would be impacted by the partialling-out of the *theta* from the raw score, as expected. The results of PCA on raw score and residuals are presented in Table 8 and Figure 1. Residual scores were further factor-analyzed through the varimax rotation method after the extraction of principal components, with a view to finding out whether locally-dependent items, possibly due to the sharing of the passages, can be identified through the factor analytic method based on Q_3. Factor loadings of 40 items on extracted factors in residuals are summarized in Table 9, and the three-dimensional visual representations of the 40 items in terms of selected factors are presented in Figures 2 through 5.

a Principal component analysis on raw scores and residuals

First, 10 principal components with eigenvalues greater than 1 were extracted for the *phi*-correlation-based factor analysis, but only 9 principal components were obtained for the *tetrachoric*-correlation-based factor analysis, as in Table 8. In the first analysis, the first principal component accounted for 20.3% percent of the total variance, and all 10 components explained approximately 47.4 % of the total variance. In the second factor analysis, on the other hand, the first principal component accounted for 30.6% percent of the total variance, which satisfies Reckase's (1979) minimal requirement (20%) but fall short of meeting Carmines and Zeller's (1979) criterion for the test to be considered to be unidimensional. All 9 components explained approximately 57.5 % of the total variance. It is known that the *tetrachoric* correlations are less likely to create the artifactual difficulty factors" than the *phi* correlations (Hulin *et al.*, 1983).

Insert Table 8 about here

As shown in Figure 1, the scree plot for the 40 items factor analysis using *tetrachoric* correlations is shaped like an elbow, which is quite often interpreted as an approximate indication of the unidimensionality of the test. The ratios of the first and second factor eigenvalues and of the difference between the first and second eigenvalues and the difference between the second and third eigenvalues (Hattie, 1985) were approximately 5.0 and 26.0, respectively, for the *phi*-correlation-based PCA and were 6.3 and 26.4 for the *tetrachoric*-correlations-based PCA. Because the *phi*-correlation-based factor analysis also produced an elbow-shaped scree plot very similar to Figure 1, only the scree plot for the *tetrachoric*-correlation-based factor is reported here. Interestingly enough, when the *tetrachoric* correlations were used in PCA, considerably more variance was explained by the first dominant, and thus the test seemed to be more unidimensional. However, when the ratios of the difference between the first and second eigenvalue and the difference between the second and third eigenvalues were almost equal for both analyses.

Insert Figure 1 about here

Second, 17 principal components with their eigenvalues greater than 1 were extracted from a matrix of interitem correlations of residual scores (Q_3). As expected, the first dominant component disappeared, after the theta is partialled out, as shown in Table 8 and Figure 1). Only the minor factors with the eigenvalues ranging from 1.91 to 1.01 were obtained. It remains to be seen whether most of these minor factors could be associated with passage-related local dependence.

b Varimax rotations on extracted factors from residuals scores.

Multiple factor analyses were conducted by reducing the number of factors to be extracted from 17 to 8 one by one in order to explore how many factors would provide the most interpretable best solution. The principal component extraction and the varimax rotation methods were used in the analysis. After the outputs from each rotation were reviewed, the 10-factor solution was selected for a detailed presentation here for two reasons: (1) it seemed to provide the most reasonable representations of the underlying factor structure related to passages; (2) the total number of passages in the test are 10. The loadings of each item on these 10 factors are summarized in Table 9. However, only the directions (+ or -) of relationships were indicated in the table for the factor loadings that are less than .25 in magnitude, except for a few cases in which it seems helpful to report the actual loadings in highlighting clusters of related items that loads most highly on a specific factor as a whole.

Insert Table 9 about here.

As shown in Table 9, about 7 passage-related factors were very clearly identified by the 10-factor solution, which include Passages 1, 2, 5, 6, 7, 8, and 10. For six of these seven passages, which are Passages 1, 2, 6, 7, 8, and 10, all of the four items in each of these passages had positive factor loadings on the same factor, and three of these four factor loadings were greater than .25. For Passage 5, however, all of the four items had negative loadings on the same factor, and three of them had factor loadings that were greater than .25 in magnitude. The relationships between these factors and items are also graphically shown in the four 3-D (3-dimensional) plots of item factor loadings (See Figure 2, 3, 4, and 5). Thus, three factors are discussed at a time in their relationships to the residuals of 40 items from now.

First, the first, second, and third factors seem to be closely related to Passages 1 (Composition), 6 (Insects), and 10 (Roller-Skating), respectively. Items from each of these three passages were found to load highly on each of these three factors respectively in the consistent manner, regardless of the number of factors extracted and rotated. In the 10-factor solution, for instance, Items 1, 2, 3, and 4 in Passage 1 were all positively correlated with the first factor, with the factor loadings of .16, .54, .40, and .37, respectively, and the 'recall of detail' (RD) item has the highest factor loading among them. There were two more RD items (i.e., Items 6 from Passage 2 and Item 34 from Passage 9) and two "prediction of content" (PC) items (Item 24 from Passage 6 and Item 36 from Passage 9) that load on Factor 1, but it seems reasonable to state that this factor may be best represented by the non-MT (main topic) items in Passage 1 (Composition).

In Figure 2, most of these items including the three items in Passage 1 items are seen to cluster around the left and middle part of the plane of Factors 1 and 2 in the 3-D plot.

Among several factors discussed here, Factor 2 and Passage 6 provide the most clear-cut example of the saliency of passage-related factors in residuals. In the 10-factor solution, three of the four items in Passage 6, which are Items 21, 22, 23, had factor loadings of .55, .53, and .50 on the second factor, and even Item 24 in the same passage also had a positive factor loading of .16 on the same factor. These three items are located near the top of the vertical axis of Factor 2 and most clearly differentiated from other items in the 3-D plot in Figure 2. No other items in the test had positive loadings on the second factor that are greater than .25. Only one item in Passage 3 had a negative loading greater than .25 in magnitude (-.38). Insightful enough, this passage was actually the one with the highest average within-passage Q_3 when weighted on the standard deviation (0.64/0.30), which was approximately 2.1 times bigger than its standard deviation (See Table 9).

Insert Figure 2 about here

The third factor appears to be related to both Passages 9 (Piano-Tuning) and 10 (Roller Skating) in the Arts and Sports topic area, but leaning more toward Passage 10. All eight items including Items 33 through 40 in both passages are positively correlated with the third factor, but the factor loading for the latter passage items (i.e., .63, .32, .55,

.22) are much higher than the former (*i.e.*, .04, .07, .34, .25). In addition, two RD items respectively from Passages 5 (Yogurt) and 7 (Polygraph) and one MT item from Passage 3 (Means for Trade) were also negatively correlated with this factor. Most of the items with the high loadings on this factor are located around the vertically middle and horizontally right corner of the 3-D plot in Figure 2.

Second, the fourth factor does not seem to be easily interpretable; the fifth factor seems to be closely related to the items in Passage 2 (English Vocabulary), whereas the sixth factor seems to be positively correlated with the items in Passage 7 (Polygraph) and negatively correlated with those in Passage 3 (Means for Trade). Items that load on Factor 4 are discussed later in this section along with Factors 9 and 10 that are also hard to interpret. When it comes to Factor 5, non-RD items (Items 5, 7, and 8) in Passage 2 had the factor loadings of .53, .72, and .31 on the fifth factor. These three items can be easily identified above the middle line along the vertical axis of Factor 5 in Figure 3, apart from the remaining items. Only item 6, an RD (recall of detail) item, has a very low factor loading of .06 on the fifth factor in Passage 2. Interestingly, another RD item, Item 30 in Passage 8 (Elevator), is the only item that had a factor loading greater with than .25 from other passages, which happens to be negative (-.29).

Insert Figure 3 about here.

All of the 4 items (Items 25, 26, 27, and 28) in Passage 7 (Polygraph) load on Factor 6, with the positive factor loadings of .28, .45, .58, and .14, respectively, but all the

items in Passage 3 (Means for Trade) were negatively correlated to this factor. The factor loadings of Items 10 and 12 only were greater than .25 in magnitude among the four items in Passage 3. These two items are seen in the horizontally left and vertically below-middle corner of Figure 3, but all of the four items from Passage 7 are located on the opposite side of the horizontal axis of Factor 6.

Third, the seventh factor seems to be closely related to items in Passage 5 (Yogurt), whereas the eighth factor seems to be closely associated to three of the four items in Passage 8 (Elevator). Factor 9, however, seems to be hard to interpret both in terms of passages and item types. All four items in Passage 8 are negatively correlated to Factor 7, with the factor loadings of -.40, -.57, -.55, and -.24 on this factor. The first (MT) and second items (RD) in Passage 3 (Means for Trade), the second item (RD) in Passage 4 (Dream), and the second item (RD) in Passage 8 (Elevator) also load on the seventh factor positively. Interestingly, three of these four items are "recall of details" items. However, the Passage 8 items still seem to represent this factor better. All of these four items are nicely displayed on the right side of the horizontal axis of Factor 7, apart from most of other items in Figure 4.

Insert Figure 4 about here.

On the other hand, three items (Items 29, 30 and 32) in Passage 8 (Elevator) were positively correlated with Factor 8, with the factor loadings of .34, .23, and .55 on this factor, respectively. Only Item 31, an "inference of details" (ID) item, had a low loading

of .03 on the eighth factor in this passage, and another ID item in Passage 4 (Dream) and one MT item in Passage 1 (Composition) had negative loadings on this factor, which were greater than .25 in magnitude on this factor. All of those three items from Passage 8 are located at the top of the plane of Factors 7 and 8 in Figure 4. Factor 9 does not seem to be a factor worth discussing.

Last, Factor 10 was included in the 3-D plot along with the two previous factors, Factors 4 and 9, which had also somewhat inconsistent patterns of factor loadings. It seems that Factor 10 may be hard to interpret. Table 9 shows that the first item (Item 13) in Passage 4 (Dream) loads on Factor 10 positively with the factor loading of .54, whereas the two first items (Items 1 and 25) respectively from Passages 1 (Composition) and 7 (Polygraph) load on this factor negatively, with the factor loadings of -.31 and -.33, respectively. Because all of these three items are MT items, such conflicting results would be problematic in proving any item-type-related local dependence consistent across multiple passages. However, the two PC (prediction of content) items respectively from Passages 2 (English Vocabulary) and 9 (Piano-Tuning) were found to load on this factor positively. There appears to be no clear, identifiable passages or item types that can be strongly associated with this factor.

On the other hand, it can be seen in Table 9 and Figure 5 that Factor 4 might be slightly easier to interpret than Factor 10. The fourth factor seems to be related weakly to Passages 4 (Dream), 7 (Polygraph), and 8 (Piano-Tuning). The third and fourth items (Items 15 and 16) from Passage 4 load more positively on Factor 4 than other items. In

contrast, the second and fourth items (Items 26 and 28) in Passage 7 and the first and third items (Items 29 and 31) in Passage 8 load negatively on Factor 4.

Insert Figure 5 about here.

In sum, it was demonstrated that the IRT theta might account for the first dominant component from the principal component analysis of raw scores well. It seems that the 40-item test used in this study does not depart seriously from the psychometrical unidimensionality despite the presence of other minor factors in the test, and many of these minor factors appears to be passage-related. Most importantly, the factor analysis of residuals based on Q_3 seems to identify items that are locally dependent, possibly due to the sharing of the passages among items or the location of the passages in the test.

IV. Discussion

The purpose of this study was (1) to examine empirically whether the sharing of a passage is a salient feature contributing to local item dependence or whether the item type also equally contributes to local item dependence in the EFL reading comprehension test and (2) to investigate whether local item dependence due to either of the above factors can be identified through the test dimensionality analysis. The findings with regard to these questions are summarized and discussed.

1. Passage-related versus item-type-related local dependence

The results of this study have clearly shown that the sharing of a passage is a salient feature contributing to local item dependence and that the item type does not

equally contributes to local item dependence in the EFL reading comprehension test used in this study. First, the average interitem correlations of raw scores within passages were higher than those within item types and the grand means of all item pairs. This has provided a preliminary piece of evidence that there might be some extra-dependence among items sharing the same passage. Nevertheless, such a pattern was not observed consistently at the levels of individual passages and item types; the average interitem correlations within a few passages were lower than those within some item types (*i.e.*,"recall of details") and close to the grand mean.

Second, the comparisons of average Q_3's within passages and within item types have provided convincing pieces of evidence that passages are the dominant factor contributing to local item dependence. The average within-passage interitem Q_3 was positive, whereas the average within-item-type $Q_3$was slightly negative and close to zero. This tendency was consistent across all the passages and item types; all of the 10 passages had positive average within-passage Q_3 values, whereas none of the 4 item types had positive average within-item-type Q_3 values.

Third, the results of factor analysis of residuals supported the similar conclusion. Seven of the ten passage-related factors were successfully identified in the 10-factor solution of the residual.

Even though the prevalence of passage-related over item-type-related local dependence was confirmed in the test, the negative average Q_3's within item types and the negative grand mean of Q_3's may need further explanation. Theoretically speaking, a negative Q_3 means that negative local dependence may exist between the two items and

that an examinee's performance on one item would negatively impacts the performance on another item. Intuitively, this negative local dependence does not make any sense for the same type of items that are sharing the same item format and supposed to tap into the similar subskills.

Two possible interpretations for the negative Q_3's for the item types can be offered. One reasonable explanation is that, as mentioned previously, even though there is no local dependence between the two items, the Q_3 could be slightly negative. Raw item scores are involved in the estimation of *theta (θ)* in IRT, and there could be the part-whole contamination between item scores and *theta (θ)* estimates (Yen, 1984). Because theta-predicted scores, residuals, and Q_3's are computed without this part-whole contamination being corrected, Q_3's may tend to become negative, even when there is no local item dependence. From this perspective, it can be said that slightly negative Q_3's within item types are an indication of no local dependence related to item types.

The second possible reason is that passage-related local dependence might have some impact on Q_3's between passages. Since the impact of passage-related local is sizeable and consistent throughout the test, it may be conceivable that a Q_3between the two items positioned in the two passages containing different content might become negative, because each item is already positively associated with one of the two passages within which they are respectively positioned in, even after the *theta (θ)* is partialled out. In this perspective, the negative Q_3 can be considered to be an indication that the passage-related local dependence is much stronger than the item-item-type-related local dependence, regardless of how much item-type-related local dependence is actually

present in the test. It is also possible that the negative Q_3's within the item types resulted from the combination of both elements.

In this study, passage-related local dependence was tested somewhat against an extreme case in this study. Only four items were used in each passage, and the same wordings were used for the item stems within the same item type so that the items can be as similar as possible in the format. In many reading comprehension tests, however, there are usually a greater variety of item types used in the same test. Besides, the current tendency is to use a larger number of items than 4 per passage, as in the paper-and-pencil TOEFL that use about 10 items per passage. As the number of questions per passage increases, items are likely to measure increasingly more redundant information from the text and become more dependent. Thus, it may be possible that local dependence within passages might be more salient under the circumstances in which a greater number of items per passage are used.

2. Local dependence and test dimensionality

The grand mean of Q_3's for all item pairs was of a negative value closer to zero, indicating that the unidimensionality assumption was seriously violated in the test as a whole. The principal component analyses of raw scores and residuals have also confirmed that there is one dominant factor that explains a large portion of the test score variance and that this dominant first factor is well explained by the latent trait (θ) modeled in the three-parameter item response theory.

There were several minor factors that influenced examinee test performance indeed. The factor analysis of residuals, however, showed that most of these minor

factors are passage-related ones, demonstrating that local item dependence is closely tied to the secondary dimensions measured by the test. However, these minor factors do not seem to be influential enough to be primary dimensions of the test. This finding also supported Schedl *et al.*'s (1996) conclusion from the TOEFL dimensionality study that passage content or passage position effects could be related to minor secondary dimensions in the EFL reading comprehension tests.

The most unexpected, but very insightful finding in terms of test dimensionality was that the positive average interitem correlations within item types were all removed, after the impact of the *theta* (θ) was partialled out from the raw scores, and this occurred consistently within the all of four item types used in this study. In a stark contrast to this, about 20% percent of the positive interritem correlations in raw scores still remain in residuals within all of the 10 passages. Even the positive correlations between different types of items disappeared, when the latent trait (θ) measured by the IRT was partialled out. This may mean that different reading subskills tapped by the same or different item types all share a common factor that fully explains the positive correlations among these items.

Nonetheless, this does not mean that the existence of such psychologically distinct subskills cannot be identified statistically. In fact, the average interitem correlation of raw scores within some item types (*e.g.*, recall of details) were higher than the grand mean of all item pairs and than the average interitem correlations within some passages, indicating that examinee performance on items within the same item type are more highly correlated than normally expected. The crucial point is, however, that the positive

correlations among these items all disappeared in Q_3, even though the positive correlations remained within those passages that had the lower raw score correlation than these item types. It implies that the unidimensional IRT theta can account for nearly all of the positive correlations among the same type of items, but cannot explain all of those among the items within the same passage.

This finding has provided support for Reckase *et al.*'s (1988) position that the unidimensional IRT models can be used, even when items clearly required more than one cognitive skills for their solutions. This is also related to Reckase's (1994) clarification that the issue of dimensionality in IRT refers only to a matrix of item response data that are the results of the interaction between examinees and test items, not to the items themselves nor to the examinees. In the same vein, Henning's (1992) distinction between "psychological unidimensionality" and "psychometric unidimensionality" has obtained an empirical validation from this study. He has suggested that it may be possible to satisfy the statistical unidimensionality constraints of some psychometric models, even when the underlying psychological trait being measured is by nature multidimensional. McNamara (1996) has also concurred that separate psychological traits could constitute a single measurement dimension, as long as they are correlated traits.

As a matter of fact, it has long been understood among language testers that the problem of reading comprehension measurement is that the "reading comprehension" construct cannot be observed directly (Kenyon, 1995). The tasks examinees perform on reading comprehension tests also might not be a direct measure of the actual reading comprehension, but of some precursory skills of reading comprehension (Anderson,

1972; Dunkel *et al.*, 1993). Information on examinees' reading comprehension should be inferred from their performance on these tasks. In this sense, it may not possibly be that surprising to find that different reading subskills tapped by these item types seem to constitute, or become part of, a unidimensional construct of reading comprehension measured by the IRT, which is basically a latent trait underlying the item responses.

V. Suggestions for Further Research

First, only four types of items were used, which are main topic inference, literal recall of details, inference of details, and prediction. A greater variety of item types could be used together with the passages of various lengths in future studies. Actually, some researchers (Hunt, 1957;Thorndike, 1971) reported that word knowledge can be distinguished from other paragraph comprehension skills.

Second, the author's subjective judgment was employed to decide the factor solution in the residual factor analysis. Confirmatory factor analysis could also be conducted in the framework of the structural equation modeling approach to come up with the most parsimonious model to explain the factor structure of the residuals.

VI. Conclusion

This study has demonstrated the evidence of a sizeable amount of passage-related local dependence in the EFL reading comprehension test and provided support for using passages as a major item grouping principle for the *testlet approach* or a *testlet response model*. Item types used in this study do not contribute to local item dependence, but different reading subskills tapped by these item types may constitute, or become part of, a unidimensional construct of reading comprehension ability measured by the IRT.

VII. References

Ackerman, T. A. 1987, April: *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Alderson, C. 1990a: Testing reading comprehension skills (part one). *Reading in a Foreign Language, 6*, 425-437.

Alderson, C. 1990b: Testing reading comprehension skills (part two). *Reading in a Foreign Language*, 7, 463-503.

Alderson, C. 1995: Response to Lumley. *Language Testing, 12*, 12-125.

Alderson, C. and **Lukmani, Y.** 1989: Cognition and levels of comprehension as embodied in test questions. *Reading in a Foreign Language, 5*, 253-270.

Alderson, J. C. and **Urquhart, A. H.** 1985: The effects of academic discipline on their performance on ESP reading tests, *Language Testing, 2*, 192-204.

Anderson, R. C. 1972: How to construct achievement tests to assess comprehension. *Review of Educational Research, 42*, 145-170.

Andrich, D. 1985: A latent trait model for items with response dependencies: Implications for test construction and analysis. In Embretson, S., editor, *Test design: Contributions from psychology, education, and psychometrics*. New York: Academic Press, 245-273.

Bachman, L. F. 1990: *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Barrett, T. 1976: Taxonomy of reading comprehension. In Smith, R. and Barrett, T. C., editors, *Teaching reading in the middle grades* (2nd edition). Reading, MA: Addison-Wesley, 61-86.

Bloom, B. S. 1956, editor: *Taxonomy of educational objectives Book 1: Cognitive domain*. London, UK: Longman.

Bradlow, E. T., Wainer, H. and **Wang, X.** 1998: *A Bayesian random effects model for testlets* (ETS research report RR-98-3). Princeton, NJ: ETS.

Brown, J. D. 1982: *Testing EFL reading comprehension in engineering English.* Unpublished Ph. D. dissertation. University of California at Los Angeles.

Buck, G. 1994: The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, *11*, 1994.

Carrell, P. 1987: Content and formal schemata in ESL reading. *TESOL Quarterly, 21*, 461-481.

Carrell, P. and **Eisterhold, J.** 1983: Schema theory and ESL reading pedagogy. *TESOL Quarterly 17*, 553-73.

Carmines, E. G. and **Zeller, R. A**. 1979: Reliability and validity of assessment. Beverly Hills, CA: Sage.

Carroll, J. B. 1969: From comprehension to inference. In Douglas, M. P., editor, *Thirty-third yearbook, Claremont Reading Conference*. Claremont, CA: Claremont Graduate School, 39-44.

Chen, W. 1996: *IRTNEW manual*. Retrieved August 28, 1997 from the World Wide Web: http://www.unc.edu/~dthissen/dl/IRTNEW.ZIP.

Chen, W., and **Thissen, D.** 1997: Local dependence indices for item pairs using item response theory. *Educational and Behavioral Statistics, 22, 265-289.*

Clapham, C. 1996: *The development of IELTS: A study of the effects of background knowledge on reading comprehension*. Cambridge, UK: Cambridge University Press.

Cureton, E. E. 1965: Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement, 25*, 327-346.

Davis, F. B. 1944: Fundamental factors of comprehension in reading. *Psychometika, 9*, 185-197.

Davis, F. B. 1968: Research in comprehension. *Reading Research Quarterly, 3*, 499-545.

Du, Z. 1998: *Modeling conditional item dependencies with a three-parameter logistic teslet model.* Unpublished doctoral dissertation. Columbia University, New York.

Dunkel, P., Henning, G., and **Chaudron, C.** 1993: The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal, 77*, 180-191.

Erickson, M. and **Molloy, J.** 1983: ESP test development for engineering students. In Oller, J. W., editor, *Issues in language testing research*. Rowley, MA: Newbury House.

Ferrera, S., Huynh, H. and **Michaels, H.** 1999: Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Assessment, 36*, 119-140.

Gray, W. S. 1960: The major aspects of reading. In Robinson, H., editor, *Sequential development of reading abilities* (Supplementary educational monograph No. 90). Chicago: University of Chicago Press.

Hale, G. H. 1988: *The interaction of student major-field group and text content in TOEFL reading comprehension* (Research Report. No. 88-1). Princeton, NJ: ETS.

Hambleton, R. K. and **Swaminathan, H.** 1985: *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Harris, D. P. 1969: *Testing English as a second language*. New York: McGraw-Hill.

Hattie, J. 1985: Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Henning, G. 1989: Meaning and implications of the principle of local independence. *Language Testing, 6*, 95-108.

Henning, G. 1990: National issues in individual assessment: The consideration of specialization bias in university language screening tests. In de Jong, J. H. L. and Stevensen, D. K., editors, *Individualizing the assessment of language abilities*. Clevedon, England: Multilingual Matters, 38-50.

Henning, G. 1992: Dimensionality and construct validity of language tests. *Language Testing, 9*, 1-11.

Hulin, C. L., Drasgow, F. and **Parsons, C. K.** 1983: *Item response theory: Application to psychological measurement*. Homewood, IN: Dow Jones-Irwin.

Hunt, L. C., Jr. 1957: Can we measure specific factors associated with reading comprehension? *Journal of Educational Research, 51*, 161-71.

Johnston, P. H. 1983: *Reading comprehension assessment: A cognitive basis.* Newark, DE: International Reading Association.

Kenyon, D. M. 1995: *Linking multiple-choice test scores to verbally-defined proficiency levels: an application to Chinese reading proficiency.* Unpublished Ph. D. Dissertation, University of Maryland, College Park.

Kingston, N. M. and **Dorans, N. J.** 1982: *The feasibility of using item response theory as a psychometric model for GRE Aptitude Test* (ETS Research Report 82-12). Princeton, NJ: Educational Testing Service.

Lee, Y. 1998: *Examining the suitability of an IRT-based testlet approach to the construction and analysis of passage-based items in an EFL reading comprehension test in the Korean High School Context.* Unpublished doctoral dissertation. The Pennsylvania State University, University Park, PA.

Little, R. J. A. and **Rubin, D. B.** 1987: *Statistical analysis with missing data.* New York: Wiley.

Lord, F. M. and **Novick, M. R.** 1968: *Statistical theories of mental test scores.* Menlo Park, CA: Addison-Wesley.

Lumley, T. 1993: The notion of subskills in reading comprehension tests: An EAP example. *Language Testing, 10*, 211-234.

Lumley, T. 1995: Response to Alderson. *Language Testing, 12*, 125-129.

McNamara, T. 1996: *Measuring second language performance.* New York: Addison Wesley Longman.

Mossenson, L., Hill, P. and **Masters, G.** 1987: *TORCH: Tests of reading comprehension* (manual). Hawthorn, Victoria: Australian Council for Educational Research.

Munby, J. L. 1978: *Communicative syllabus design*. Cambridge, UK: Cambridge University Press.

Pearson, P. D. and **Johnson, D. D.** 1978: *Teaching reading comprehension*. New York: Holt, Rinehart, and Winston.

Rechase, M. D. 1994: What is the "correct" dimensionality for a set of item response data? In Laveault, D., Zumbo, B. D., Gessaroli, M. E., and Boss, M. W., editors, *Modern theories of measurement: Problems and issues*. Ottawa, Canada: Edumetrics Research Groups, University of Ottawa, 87-92.

Reckase, M. D., Ackerman, T. A. and **Carlson, J. E.** 1988: Building unidimensional tests using multidimensional items. *Journal of Educational Measurement, 25*, 193-203.

Rosenbaum, P. R. 1988: Item bundles. *Psychometrika, 53*, 349-359

Rumelhart, D. E. 1980. Schemata: the building blocks of cognition. In Spiro, R. J., Bruce, B. C. and Brewer, W. F., editors, *Theoretical issues in reading comprehension*, Hillsdale, NJ: Erlbaum, - .

Savignon, S. J. 1983: *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.

Schedl, M., Gordon, A., Carey, P. and **Tang, K. L.** 1996: *An analysis of the dimensionality of TOEFL reading comprehension items* (TOEFL research report 53). Princeton, NJ: ETS.

Sireci, S., Thissen, D. and **Wainer, H.** 1991: On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Spray, J. A. and **Ackerman, T.** 1987: *The effects of item response dependency on trait or ability dimensionality* (ACT Research Report Series 87-10). Iowa City, IA: ACT.

Steinberg, L. and **Thissen, D.** 1996: Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81-97.

Stout, W. F., Douglas, J., Junker, B. and **Roussos, L.** 1993: *DIMTEST manual.* Unpublished manuscript, University Illinois, Urbana-Champaign.

Suen, H. K. 1990: *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

Thorndike, R. L. 1971: *Reading as reasoning.* Address delivered to Division 15, American Psychological Association, Washington, DC.

Thissen, D. 1991: *MULTILOG user's guide.* Chicago: Scientific Software.

Thissen, D. and **Chen, W.** 1993: *Item response theory and local dependency: An interim report* (Research Memorandum 93-3). Chapel Hill, NC: L. L. Thurston Laboratory, University of North Carolina at Chapel Hill. .

Thissen, D., Steinberg, L. and **Mooney J. A.** 1989: Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26*, 247-260.

van den Wollenberg, A. L. 1982: Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

Wainer, H., Bradlow, E. T. and **Du, Z.** 1999: Testlet response theory: An analog for the 3-PL. In van der Linden, W. J. and Glas, C. A. W., editors, *Computerized Adaptive Testing*. Boston, MA: Kluwer-Nijhoff.

Wainer, H. and **Thissen, D.** 1996: How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22-29.

Weiss, D. J. and **Yoes, M. E.** 1991: Item response theory. In Hambleton, R. K. and Zaal, J. N., editors, *Advances in educational and psychological testing: Theory and applications*. Norwell, MA: Kluwer-Academic, 69-95.

Wilson, M. and **Adams, R.** 1995: Rasch models for item bundles. *Psychometrika, 60*, 181-198.

Wilson, M. and **Iventosch, L.** 1988: Using the partial credit model to investigate responses to structured subtests. *Applied Measurement in Education, 1*, 319-334.

Yen, W. M. 1981: Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. M. 1984: Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. 1993: Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Table 1
Relationships between Questions and Text Information in 4 Item Types

Item Types	Relationship between questions and text info.	Response Modes
Recall of Details	textually explicit	Multiple-Choice
Main Topic Inference	textually implicit	Multiple-Choice
Inference of Details	textually implicit	Multiple-Choice
Prediction of Content	scriptally implicit	Multiple-Choice

Table 2
Item Types, Titles, Topic Areas, and the Number of Words for 10 Passages

Passages	Items 1 2 3 4	Titles	Topic Areas	Word Count
P01	MT RD ID PC	Composition	Humanity	148
P02	MT RD ID PC	English vocabulary	Humanity	194
P03	MT RD ID PC	Means for trade	Social Science	181
P04	MT RD ID PC	Dream	Social Sciences	200
P05	MT RD ID PC	Yogurt	Natural Science	166
P06	MT RD ID PC	Insects	Natural Science	181
P07	MT RD ID PC	Polygraph	Engineering	203
P08	MT RD ID PC	Elevator	Engineering	188
P09	MT RD ID PC	Piano Tuning	Arts & Sports	178
P10	MT RD ID PC	Roller skating	Art & Sports	193

Note: "MT", "RD", "ID", and "PC" are 4 item types used in each passage, which stands for "main topic inference", "recall of details", "inference of details", and "prediction of content" items, respectively.

Table 3
The Reliability Estimates for Dichotomous and Polytomous Scoring
(N=1072: Listwise Deleted; N=1857: Zero-Replacement)

Scoring Unit	Scoring Scheme	No. of Items	No. of Cases	Mean Score	SD	Average Inter-Item Corre.	Cronbach- α (KR-20)
Item	Dichotomous	40	1072	20.4	8.58	.173	.893
Passage	Polytomous	10	1072	20.4	8.58	.400	.869
Item Type	Polytomous	4	1072	20.4	8.58	.679	.893

Table 4
Average Interitem Correlations of Raw Scores Within Passages and Item Types

Corr. Type	Missing Cases	No. of Cases	Within Passages	Between Passages SameType	Diff.Type	Total	Grand Mean
Phi	Listwise	1072	.229	.173 (.058)	.167	.168	.173 (.058)
Phi	Pairwise	1576 *	.215	.166 (.050)	.161	.162	.166 (.050)
Tetra	Listwise	1072	.363	.275 (.098)	.267	.269	.276 (.097)

Note: 1. "*" indicates the average number of cases over all item pairs.
2. Numbers in parentheses are the differences between the average within-passage correlation and the average correlations to the left of the parentheses.

Table 5
Average Interitem Correlations of Raw Scores Within Each Passage and Each Item Type

(a) Within Passages

Comp. Method	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	PMean
Phi-L											
Mean	.178	.176	.286	.242	.284	.264	.224	.253	.182	.193	.229
SD	.065	.141	.090	.050	.144	.051	.077	.053	.071	.109	.095
Phi-P											
Mean	.162	.161	.276	.220	.280	.257	.204	.231	.177	.174	.215
SD	.047	.142	.097	.041	.148	.051	.069	.059	.053	.108	.095
Tet-L											
Mean	.290	.284	.478	.400	.484	.432	.376	.410	.309	.335	.380
SD	.101	.224	.147	.084	.223	.081	.136	.086	.125	.175	.153

(a) Within Item Types

Comp. Method	ITP1	ITP2	ITP3	ITP4	ITPMean
Phi-L					
Mean	.190	.199	.174	.127	.173
SD	.054	.098	.049	.072	.076
Phi-P					
Mean	.184	.195	.168	.118	.166
SD	.053	.104	.045	.073	.078
Tet-L					
Mean	.310	.328	.280	.211	.282
SD	.082	.161	.080	.119	.123

Note: "Phi-L," "Phi-P," and "Tet-L" refer to the *phi* correlations when cases with missing values were deleted listwise (N=1072) and pairwise (*N=1576), and the *tetrachoric* correlations when cases were with missing values were deleted listwise, respectively.

Table 6
Average Interitem Correlations of Residuals (Q_3) Within Passages and Item Types

Theta-Estimation			Q_3 Computation		Within Passages	Between Passages			Grand Mean
Esti. Meth.	Msng Data	No. of Cases	Msng Data	No. of Cases		Same Type	Diff. Type	Total	
MLE	List	1072	NA	1072	.051	-.019 (.069)	-.023	-.022	-.016 (.067)
MAP	List	1072	NA	1072	.051	-.019 (.069)	-.023	-.022	-.016 (.067)
MLE	9*	1797	List	1041	.051	-.020 (.071)	-.024	-.023	-.018 (.068)
MLE	9*	1797	Pair	1523**	.044	-.020 (.064)	-.023	-.023	-.018 (.061)
MAP	9*	1857	List	1072	.052	-.018 (.071)	-.023	-.022	-.016 (.068)
MAP	9*	1857	Pair	1576**	.042	-.022 (.064)	-.025	-.025	-.019 (.061)

Note: 1. MLE and MAP refer to "Maximum Likelihood Estimates" and "Maximum A Posteriori" estimation methods, respectively.
2. "List," "Pair," and "NA" refer to "listwise deletion," "pairwise deletion," and "not applicable."
3. "*" indicates that "9" was used as a missing data code in MULTILOG.
4. "**" indicates the average number of cases over all item pairs.

Table 7

Average Interitem Correlations of Residuals (Q_3)
Within Each Passage and Each Item Type

(a) Within Passages

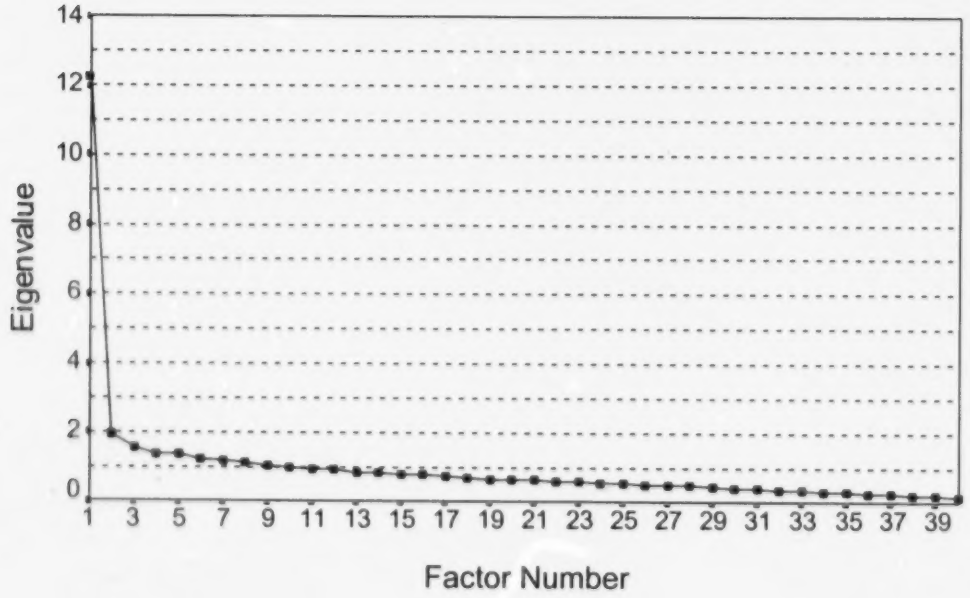
Comp. Method	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	PMean
Lw-0											
Mean	.076	.049	.038	.023	.034	.064	.056	.068	.025	.072	.051
SD	.043	.088	.046	.039	.040	.030	.032	.036	.062	.062	.050
9-Lw											
Mean	.072	.055	.045	.023	.026	.060	.058	.072	.023	.070	.051
SD	.046	.086	.049	.040	.032	.027	.033	.031	.060	.060	.050
9-Pw											
Mean	.064	.046	.041	.016	.025	.062	.046	.050	.026	.064	.044
SD	.024	.069	.046	.039	.029	.027	.021	.042	.045	.041	.038

(b) Within Item Types

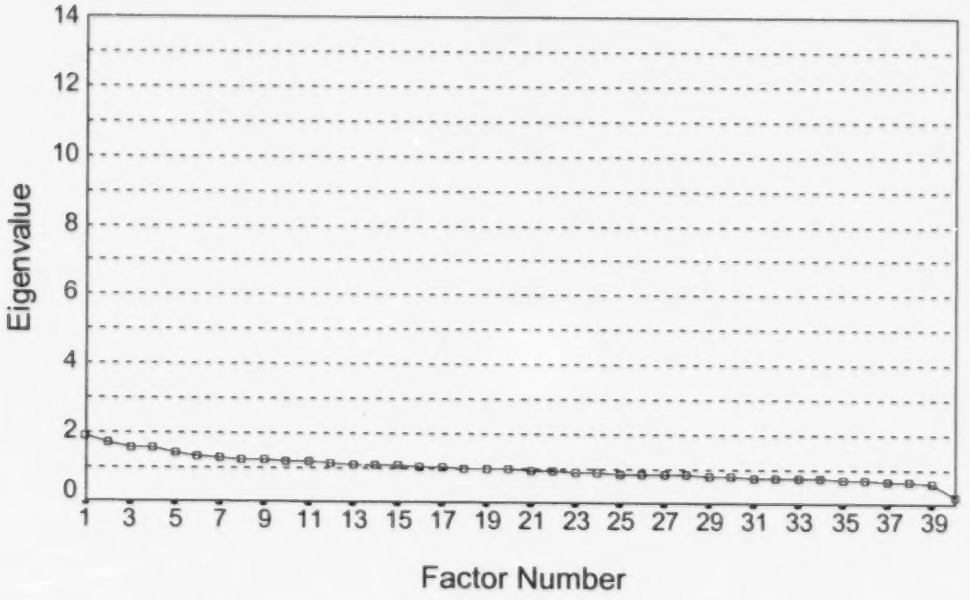
Comp. Method	ITP1	ITP2	ITP3	ITP4	ITPMean
Lw-0					
Mean	-.029	-.018	-.021	-.007	-.019
SD	.042	.064	.046	.043	.050
9-Lw					
Mean	-.031	-.021	-.021	-.009	-.020
SD	.043	.062	.047	.041	.049
9-Pw					
Mean	-.028	-.022	-.021	-.010	-.020
SD	.038	.058	.038	.038	.044

Note: 1. "Lw-0," "9-Lw," and "9-Pw" refer to the MLE (Maximum Likelihood Estimate) method for the first data set (N=1072), the MLE/listwise deletion method for the second data set (N=1781), and the MLE/pairwise deletion method for the second data set (N=1857), respectively.

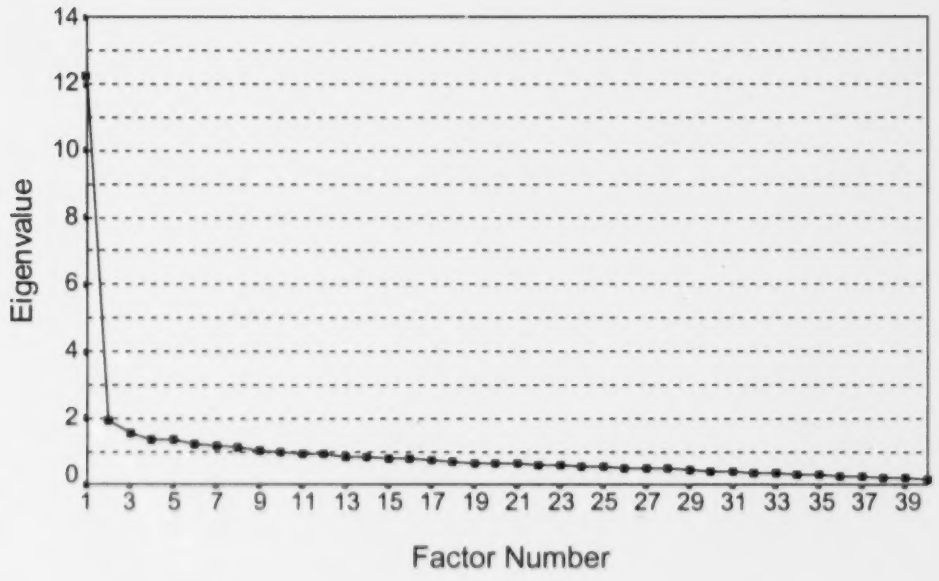
2. Cases with missing values were deleted listwise in the first data set (N=1072) before theta estimation and coded as 9 in the second data set (N=1781), and cases with no *theta* estimates were eliminated from the analysis..



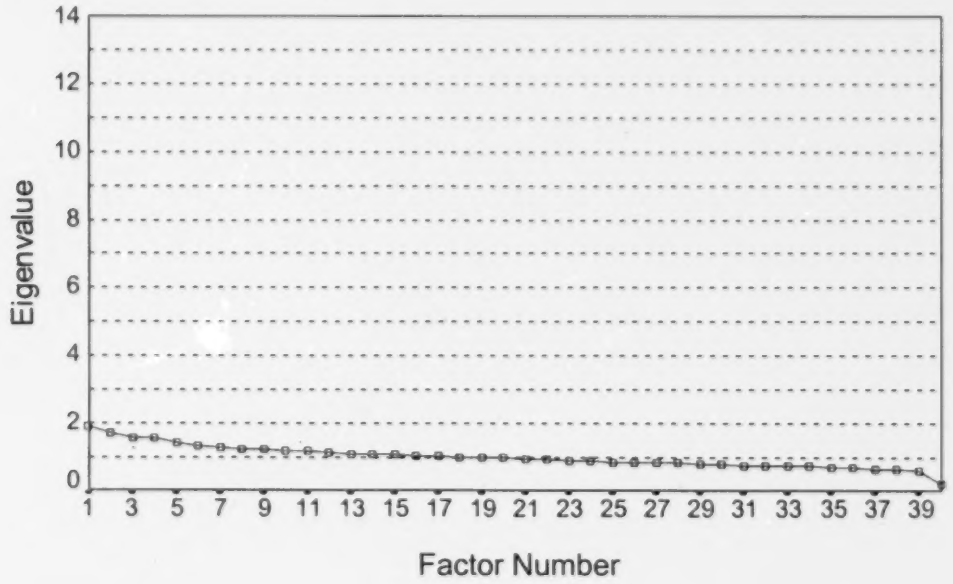
(a) Raw Scores (Tetrachoric)



(b) Residuals (Pearson)



(a) Raw Scores (Tetrachoric)



(b) Residuals (Pearson)

Figure 1 Factor Scree Plots from Principal Component Analysis of 40 Items

Table 9
Factor Loadings of 40 Items on 10 Major Factors in the Residuals
(N=1072)

Passages	Items	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
	1	(.16)	-	+	-	+	-	+	-.50	-	-.31
1	2	.54	-	-	-	+	+	-	-	-	-
	3	.40	-	-	.25	+	-	+	-	+	-
	4	.37	-	+	+	-	-	-	+	+	+
	5	-	-	+	+	.53	-	+	-	-	-
2	6	.43	+	-	+	(.06)	+	+	+	+	-
	7	+	-	-	+	.72	+	+	+	+	+
	8	+	+	+	-.32	.31	+	+	-	+	.25
	9	-	+	-.25	+	+	(-.12)	.38	-	-	+
3	10	-	-	-	-	-	-. 59	.28	-	+	+
	11	-	-	+	+	-	(-.03)	+	+	+	-
	12	-	-.38	-	+	-	-.33	+	-	-	+
	13	-	-	-	+	-	+	+	+	+	.54
4	14	+	+	-	+	-	-	.25	+	.30	+
	15	+	-	-	.34	-	-	-	-.33	+	+
	16	-	-	-	.48	+	+	-	+	-	+
	17	+	+	-	+	-	-	-.40	+	+	+
5	18	-	+	-.25	+	+	+	-.57	+	-	-
	19	-	+	+	+	-	-	-.55	-	-	+
	20	+	-	+	-	-	-.29	(-.24)	+	+	-
	21	-	.55	-	-	-	-	-	-	-	-
6	22	-	.53	-	+	+	-	-	+	+	-
	23	-	.50	+	+	-	+	-	-	-	+
	24	.33	(.16)	-	+	-	-	-	+	+	-
	25	-	+	+	-	-	.28	+	-	.34	-.33
7	26	-	-	-	-.32	+	.45	+	-	-	-.25
	27	-	-	-	+	-	.58	+	-	+	+
	28	-	-	+	-.42	+	(.14)	-	+	+	+
	29	+	-	-	-.27	-	+	-	.34	-	-
8	30	-	-	-	+	-.29	+	.27	.23	-.44	-
	31	-	-	-	-.52	-	-	+	(.03)	-	-
	32	+	-	+	-	+	-	+	.55	-	-
	33	-	+	(.04)	+	+	-	+	+	-.26	+
9	34	.39	-	(.07)	-.25	+	-	+	+	-	-
	35	+	+	.34	-	-	+	+	-	+	+
	36	.28	-	.25	-	-	+	+	+	-	.39
	37	+	+	.63	-	+	-	+	+	-	+
10	38	+	-	.32	+	-	+	+	+	-	-
	39	-	-	.55	+	+	-	+	+	+	-
	40	-	-	(.22)	.26	+	+	-	-	+	-

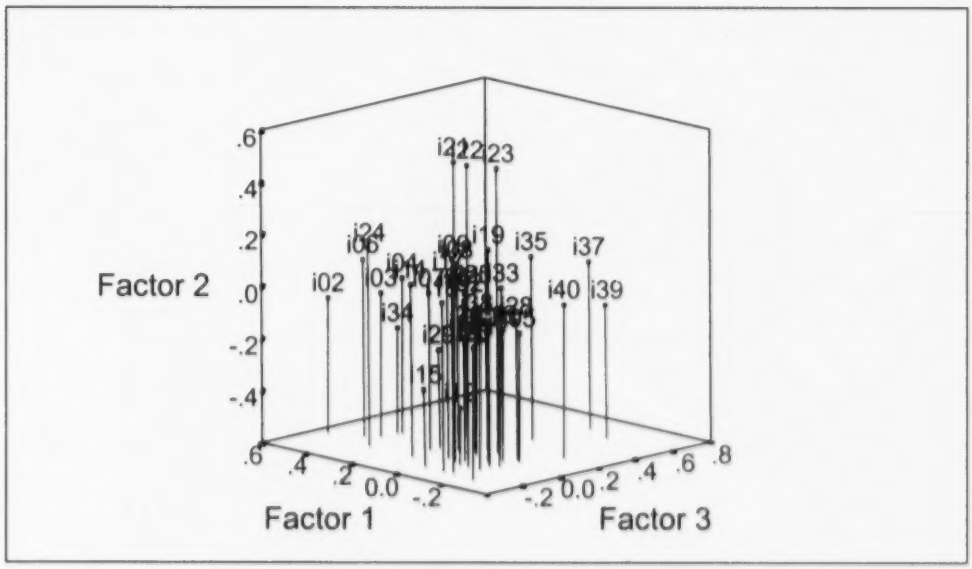


Figure 2. Three-Dimensional Plot of Loadings of 40 Items on Factors 1, 2, and 3.

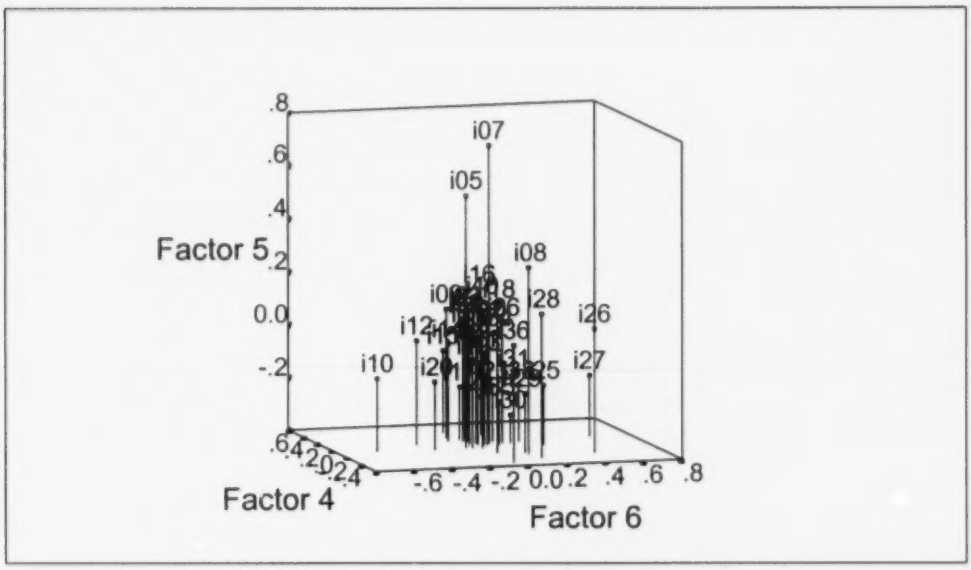


Figure 3. Three-Dimensional Plot of Item Factor Loadings on Factors 4, 5, and 6

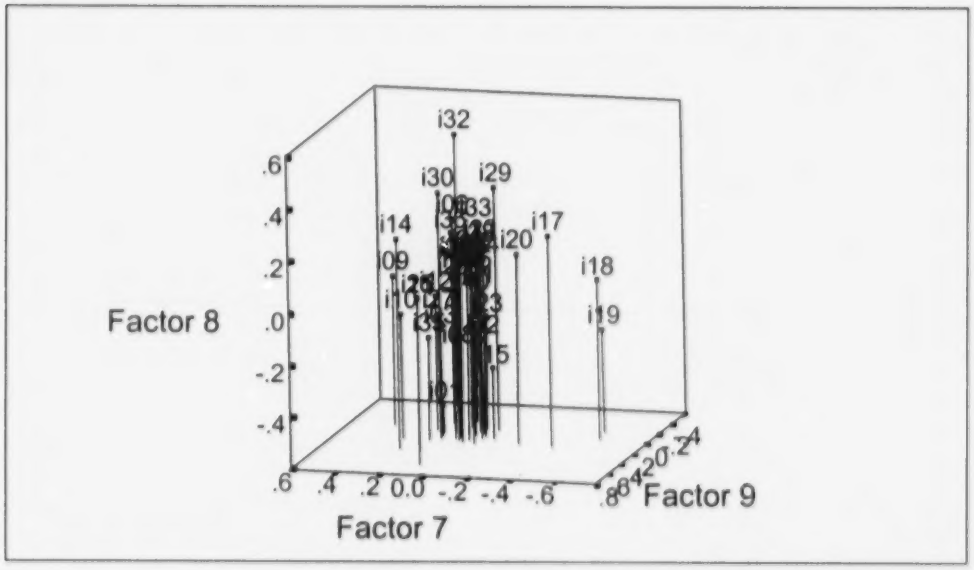


Figure 4. Three-Dimensional Plot of Loadings of 40 items on Factors 8, 7, and 9.

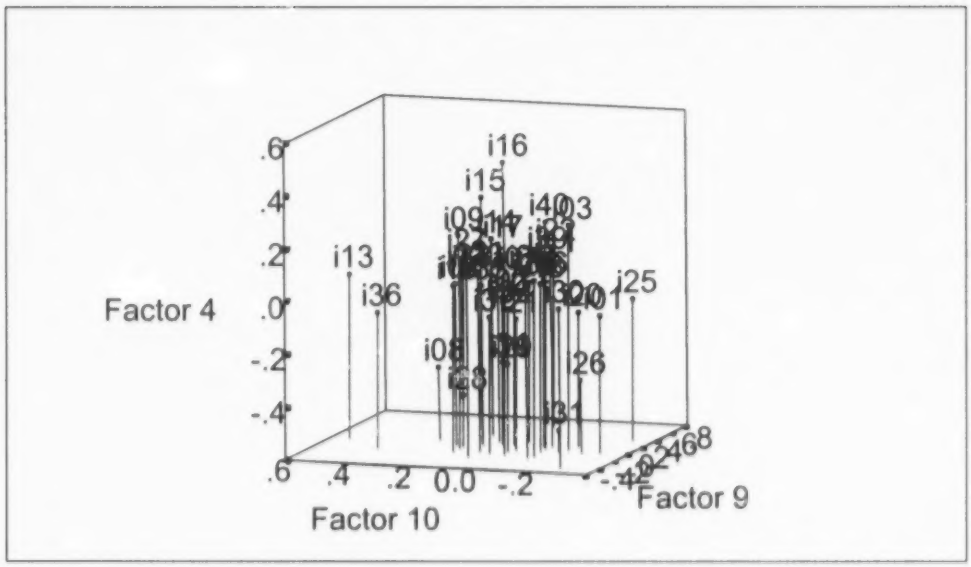


Figure 5. Three-Dimensional Plot of Loadings of 40 Items on Factors 4, 9, and 10